Understanding the Query:
THCIB and THUIS at NTCIR-10 Intent Task

Yunqing Xia\textsuperscript{1} and Sen Na\textsuperscript{2}
\textsuperscript{1} Tsinghua University
\textsuperscript{2} Canon Information Technology (Beijing) Co. Ltd.
Before we start

• Who are we?
  – **THUIS** is the research team at Intelligent Search group at Center for Speech and Language Technology, Tsinghua University
  – **THCIB** is the joint research team between THUIS and Canon Information Technology (Beijing) Co. Ltd..

• Why did we participate NTCIR INTENT task?
  – We believe intent mining is one of the most promising technologies to make the search engines smarter thus more helpful to human.
  – We view query-based intent mining as a major topic in our research group

• What is task/subtask we participated?
  – Subtopic mining: Systems are required to return a ranked list of *subtopic strings* in response to a given topic query while the top N subtopic strings should be *both relevant and diversified* as much as possible.
Outline

• The motivation
• System overview
• What make our system different?
• Evaluation
  – The submitted runs
  – Results and discussion
• Conclusion and future work
The Motivation (1/3)

- ISSUE #1: Query is usually very short

- SOLUTION #1: Applying BIGGER CONTEXT in query understanding
  - User behavior data: Query log, search engine auto-completions and suggestions
  - Search results: Title and snippet
• ISSUE #2: Subtopic surface strings are redundant

furniture for small spaces store
{furniture for small spaces market}
{furniture for small spaces wholesale}
{furniture for small spaces shop}
{furniture for small spaces center}
......

{furniture for small spaces Tokyo}
furniture for small spaces New York
{furniture for small spaces London}
{furniture for small spaces Hong Kong}
{furniture for small spaces Indonesia}
......

• SOLUTION #2: Discover the implicit intents by clustering the subtopic surface strings
  – A sense-based clustering algorithm
The Motivation (2/3)

• **ISSUE #3:** Relevance is no longer effective for intent ranking

• **SOLUTION #3:** Ranking intents considering both relevance and diversity
  – A unified intent weighting model and a subtopic selecting strategy
Outline

• The motivation
• System overview
• What make our system different?
• Evaluation
  – The submitted runs
  – Results and discussion
• Conclusion
System overview

Subtopic candidate mining (SCM)

Subtopic candidate ranking (SCR)
Outline

• About the NTCIR10 INTENT-2 task
  – Who are we?
  – Why do we participate NTCIR INTENT task?
  – Task/subtask we participated
• The motivation
• System overview
• What make our system different?
• Evaluation
  – The submitted runs
  – Results and discussion
• Conclusion and future work
What make our system different?

• Concept based
  – Wikipedia entries and related entries
  – From query analysis to expansion
  – From subtopic extraction to intent mining
  – From relevance to diversity
  – From weighting to ranking

• Discovering intent for diversification
  – Word sense induction
  – Intent induction/disambiguation
  – Entity analysis to address homogeneous exclusive subtopics
SCM: Extracting concepts from query

- Downloading the entire Wikipedia
  - Entry ==> Concept
  - Concept ==> Definition
  - Concept ➔ Related concepts

- Bi-directional maximum entry matching
- Using the multiple matches in the disambiguation page
- Using redirects when no entry is exactly matched

“battles in the civil war” ➔ “battle”, “civil war”
SCM: Query expansion

• Wikipedia
  – Synonymous entries (redirects) and the related concepts
  – Polysemous entries (disambiguation pages)

• Intent schema
  – \{concepts, prepositions, wild cards\}
  – “hobby store”: “* of hobby store”, “* at hobby store”,
    “hobby store in *”, “hobby store at *”, etc.

• Concept repositioning
  – “battles in civil war” \(\Rightarrow\) “battles civil war”, “civil war battles”

• The motivation:
  – Reforming the query so as to obtain subtopic candidates as many as possible (in query auto-completion, query suggestions, etc.)
SCM: Extracting subtopic candidates

- **Wikipedia** – general knowledge base
  - Concept definition
- **User Behavior Data** – user centric data
  - Co-occurrence
  - Search engine tools (auto-completion, query suggestion)
- **Search Results** – pseudo feedback
  - Query topics (word senses) within snippets of top N=1000 results
Wikipedia concept definition

- **Ground**
  - Kick scooter
  - Motorized scooter
  - Scooter (motorcycle)
  - Knee scooter
  - Mobility scooter
  - Eccentric-hub scooter
  - Square scooter
- **Air**
  - Douglas A-4 Skyhawk
  - Air Scooter
- **Water**
  - Underwater scooter
  - Water scooter
  - Ice boat
- **People**
  - Scooter Braun
  - Lloyd L. Burke (nicknamed "Scooter")
  - Dill Stokes (nicknamed "Scooter")
- **Fictional characters**
  - Scooter (comics)
  - Scooter (Coronation Street)
  - Scooter (Gobots)
  - Scooter (Muppet)
  - Scooter (SpongeBob SquarePants)
  - Scooter (talking baseball)
  - Scooter: Secret Agent
User Behavior Data

• The user search log (e.g., ClueWeb09)

• Tools of commercial search engines based on user behavior data
  – Auto-completion
  – Query suggestion

• With expanded queries based on concepts
Search results

• Search with concept as a whole keyword
  – In query <battles in the civil war>, <“civil war”> is one keyword WORD
  – In Web pages, <“civil war”> is one keyword WORD (‘war’ must immediately follow ‘civil’)

• Induce aspects of the query using WSI (word sense induction) technique
  – LDA + keyword extraction
  – Labeled LDA
  – Sense based LDA: a sense based clustering algorithm

SCR: Re-calculating the relevance score

• Replacing bag of word with bag of Wikipedia concepts
  – BM25 again.

• Incorporating source score

\[ p_t = w_{ST}(t) + w_{SC}(t) \]

• \( w_{st}(t) \): Relevance score of the subtopic candidates
• \( w_{sc}(t) \): Importance score (empirical) of the source where the subtopic comes from.
SCR: Discovering intents

• Clustering subtopics candidates with Affinity Propagation (AP) algorithm
  – Calculating subtopic similarity with VSM-based cosine similarity
  – Extraction concept-based VSM features from snippets of the top 50 search results with subtopic string as a query.
  – Choosing mean of the similarity matrix as clustering preference value

• The revised version
  – Choosing mean of the subtopic importance value (=relevance + resource weight)
SCR: Weighting the intents

- A simple sum equation

\[ w_{IN} = \sum_{i=1}^{N} [w_{ST}(t_i) + w_{SC}(t_i)] \]
SCR: Entity analysis

- Homogenous exclusive entities are found many in subtopic candidates
  - “furniture for small spaces New York”
  - “furniture for small spaces Los Angeles”

- Freebase - a global resource of ontology
  - It provides HTTP API for data retrieval
  - The whole dump data can be downloaded from Web

- Judgment of homogenous exclusive entities
  - Sharing the same immediate father node!
  - “City/Town/Village”
SCR: Selecting for ranking
Outline

• The motivation
• System overview
• What make our system different?
• Evaluation
  – The submitted runs
  – Results and discussion
• Conclusion and future work
The submitted runs

- We submitted 5 runs for English task

<table>
<thead>
<tr>
<th>RUN ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>THCIB-S-E-2A</td>
<td>THCIB-S-E-1A + SCM (2.Query expansion)</td>
</tr>
<tr>
<td>THCIB-S-E-3A</td>
<td>THCIB-S-E-2A + SCR (4.Entity analysis)</td>
</tr>
<tr>
<td>THCIB-S-E-5A</td>
<td>THCIB-S-E-4A + SCR (3.Intent mining with revised AP)</td>
</tr>
</tbody>
</table>

- We submitted 4 runs for Chinese task
  - No Freebase in Chinese (Run 3 in English is not planned for Chinese task).
Results and discussion – Performance

• Rank:
  - Run 2 > Run 1 > Run 3 > Run 5 > Run 4

• Observations
  - Concept-based query expansion is useful in subtopic mining (Run 2 vs. 1)
  - Entity analysis is not appropriately used (Run 4)
  - Performance of intent discovery can be improved (Run 3)
  - Intent-based subtopic weighting model can be improved (Run 3)

• Performance in Chinese task is similar

<table>
<thead>
<tr>
<th>cut-off</th>
<th>run name</th>
<th>I-rec</th>
<th>D-nDCG</th>
<th>D#-nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>@10</td>
<td>THCIB-S-E-1A</td>
<td>0.3785</td>
<td>0.3384</td>
<td>0.3584</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-2A</td>
<td>0.3797</td>
<td>0.3499</td>
<td>0.3648</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-3A</td>
<td>0.3681</td>
<td>0.3383</td>
<td>0.3532</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-4A</td>
<td>0.3502</td>
<td>0.3323</td>
<td>0.3413</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-5A</td>
<td>0.3662</td>
<td>0.3215</td>
<td>0.3438</td>
</tr>
<tr>
<td>@20</td>
<td>THCIB-S-E-1A</td>
<td>0.5769</td>
<td>0.3274</td>
<td>0.4522</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-2A</td>
<td>0.5899</td>
<td>0.3406</td>
<td>0.4653</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-3A</td>
<td>0.5544</td>
<td>0.3251</td>
<td>0.4397</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-4A</td>
<td>0.477</td>
<td>0.2784</td>
<td>0.3777</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-5A</td>
<td>0.5395</td>
<td>0.304</td>
<td>0.4218</td>
</tr>
<tr>
<td>@30</td>
<td>THCIB-S-E-1A</td>
<td>0.693</td>
<td>0.3177</td>
<td>0.5054</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-2A</td>
<td>0.6743</td>
<td>0.3284</td>
<td>0.5014</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-3A</td>
<td>0.6486</td>
<td>0.3244</td>
<td>0.4865</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-4A</td>
<td>0.5855</td>
<td>0.2691</td>
<td>0.4273</td>
</tr>
<tr>
<td></td>
<td>THCIB-S-E-5A</td>
<td>0.6339</td>
<td>0.2986</td>
<td>0.4662</td>
</tr>
</tbody>
</table>
Results and discussion – Per-topic analysis

• Best runs on the 50 queries
  – THCIB-S-E-1A 8
  – THCIB-S-E-2A 13
  – THCIB-S-E-3A 6
  – THCIB-S-E-4A 13
  – THCIB-S-E-5A 10
  – No run is consistently best, and each shows strength (further study is necessary)

• Query length
  – Our system is not sensitive to query length (Num. of words)
  – Other factors should be studied.
Outline

• The motivation
• System overview
• What make our system different?
• Evaluation
  – The submitted runs
  – Results and discussion
• Conclusion and future work
Conclusion and future work

• Conclusion
  • Incorporating concepts and word senses in subtopic mining and ranking brings marginal performance gain (NLP is positive to SM).
  • Subtopic ranking based on the automatically discovered intent is promising (though more work is required to improve intent quality).

• Future work
  • Deeper understanding the query: better subtopic extraction and intent discovery
  • Complexity issue: concept based indexing and retrieval
  • How about navigational and transactional query?
Acknowledgement

We thank Canon Inc. for supporting this research (No. TEMA2012).

We also thank the valuable comments from INTENT-2 organizer.
THANK YOU!
Q&A

We also welcome offline discussion by sending emails to
yqxia@tsinghua.edu.cn